Cardiovascular Transcriptomics and Epigenomics Using Next-Generation Sequencing Challenges, Progress, and Opportunities

Po-Yen Wu, *Georgia Institute of Technology*
Raghu Chandramohan, *Georgia Institute of Technology*
John Phan, *Emory University*
William Mahle, *Emory University*
J. William Gaynor, *University of Pennsylvania*
Kevin Maher, *Emory University*
May D. Wang, *Emory University*

Journal Title: Circulation: Cardiovascular Genetics
Volume: Volume 7, Number 5
Publisher: American Heart Association | 2014-10-01, Pages 701-710
Type of Work: Article | Post-print: After Peer Review
Publisher DOI: 10.1161/CIRCGENETICS.113.000129
Permanent URL: https://pid.emory.edu/ark:/25593/rr3c3

Final published version:
http://dx.doi.org/10.1161/CIRCGENETICS.113.000129

HHS Public Access

Author manuscript

Circ Cardiovasc Genet. Author manuscript; available in PMC 2016 August 14.

Published in final edited form as:

Circ Cardiovasc Genet. 2014 October ; 7(5): 701–710. doi:10.1161/CIRCGENETICS.113.000129.

Cardiovascular Transcriptomics and Epigenomics Using Next-Generation Sequencing: Challenges, Progress, and Opportunities

Po-Yen Wu, MS[1], **Raghu Chandramohan, MS**[2], **John H. Phan, PhD**[3], **William T. Mahle, MD**[4], **J. William Gaynor, MD**[5], **Kevin O. Maher, MD**[6], and **May D. Wang, PhD**[7]

[1]School of Electrical and Computer Engineering, Georgia Institute of Technology

[2]School of Biology, Georgia Institute of Technology

[3]The Wallace H Coulter Department of Biomedical Engineering, Georgia Institute of Technology & Emory University

[4]Children's Healthcare of Atlanta, Emory University School of Medicine, Atlanta, GA

[5]The Children's Hospital of Philadelphia, Philadelphia, PA

[6]Children's Healthcare of Atlanta, Atlanta, GA

[7]School of Electrical and Computer Engineering, Winship Cancer Institute, Parker H. Petit Institute for Bioengineering and Bioscience, Georgia Institute of Technology & Emory University

Introduction: Next-Generation Sequencing for Personalized Cardiovascular Disease Care

Cardiovascular disease (CVD) is the leading cause of death worldwide. Prediction and prevention of CVD, such as coronary artery disease and atherosclerosis, traditionally depend on identification of risk factors.[1, 2] These factors are effective in the general assessment of CVD risk, but are not consistent indicators for all individuals.[3] Therefore, CVD research has recently been expanded to include the identification of omic biomarkers (e.g., genomic, transcriptomic, and epigenomic) that may (1) improve our understanding of the molecular mechanisms of CVD, (2) facilitate the development of personalized CVD care, and (3) reduce CVD mortality rates by accurately identifying high-risk individuals.[4] Next-generation sequencing (NGS) is a promising technology to identify omic biomarkers. Because of its high-throughput capability in discovering novel genomic features with base-pair resolution, NGS is projected to play an increasingly important role in clinical diagnostics and personalized medicine for CVD.[5, 6]

Correspondence: May D Wang, PhD, Biomedical Engineering Department, Georgia Institute of Technology & Emory University, 313 Ferst Drive, UA Whitaker Building, Suite 4106, Atlanta, GA 30332, Tel: 404-385-2954, Fax: 404-894-4243, maywang@bme.gatech.edu.

Conflict of Interest Disclosures: None

NGS and associated bioinformatics methods have been applied to cardiovascular genomics, transcriptomics, and epigenomics. Figure 1 illustrates four NGS applications such as (A) identification of differentially expressed genes (DEG) using RNA sequencing (RNA-seq), (B) identification of protein-binding regions in the genome using chromatin immunoprecipitation sequencing (ChIP-seq), (C) identification of genetic variants in exon regions using exome sequencing, and (D) identification of genomic methylation patterns using methyl-CpG-binding domain sequencing (MBD-seq). These applications identify and quantify omic biomarkers that may be clinically viable for early disease diagnosis and effective disease treatment and management. In this article, we focus on two major applications of NGS technology: (1) RNA-seq, which has enabled researchers to characterize CVD by studying transcriptome-wide expression profiles,[7] alternative splicing patterns,[8] and miRNA regulatory networks;[9] and (2) ChIP-seq, which has enabled researchers to examine the epigenetic mechanisms of CVD by profiling the genome-wide pattern of protein-binding regions (e.g., transcription factors and enhancers) or histone modifications.[10, 11]

Although NGS provides a great opportunity for discovering potential CVD biomarkers, a few NGS bioinformatics challenges remain: First, the overwhelming NGS data volume requires huge data storage and computational resources. Second, at each data analysis step, multiple bioinformatics tools are publicly available, which makes it challenging to assemble sensible NGS bioinformatics pipelines. Third, one single NGS experiment processed by various NGS bioinformatics pipelines can result in a wide variety of information with different biological and clinical interpretation and validation. In Section 2, we will review the challenges and progress of bioinformatics in NGS data analysis with a focus on RNA-seq and ChIP-seq. In Section 3, we will present two CVD case studies to illustrate the applications of NGS bioinformatics. In Section 4, we will summarize future opportunities for CVD research using NGS.

NGS Bioinformatics for CVD

Illumina, Life Technologies/Ion Torrent, and Roche/454 are examples of commercially available NGS platforms. Among these, Illumina is the most prevalent platform that can produce millions of relatively short, fixed-length sequence “reads” in a single experiment. In this section, we will describe the NGS bioinformatics pipelines tailored to handle Illumina data. As shown in Figure 2, we will review bioinformatics methods such as sequence mapping, expression quantification, expression normalization, and DEG detection for RNA-seq data, and sequence mapping and peak calling for ChIP-seq data.

Sequence Mapping

The first step of bioinformatics pipelines for both RNA-seq and ChIP-seq is sequence mapping. It determines the genomic or transcriptomic origin of sequence reads (or “reads” in short). Sequencing mapping using Brute-force strategy requires large CPU and memory resources, where mapping millions of reads to the three billion base pairs of the human genome is extremely time-consuming. Thus, the research of sequencing mapping largely focuses on improving computational efficiency while maintaining high mapping accuracy.

Table 1 lists some mapping tools with their mapping strategies and key features. Depending on biological applications and computational resources, mapping algorithms can provide three types of alignments: (1) Un-gapped alignment (e.g., Bowtie[12]) allows only mismatches between query reads and the reference genome to keep the computational cost low. However, for some applications (e.g., mapping of RNA-seq data to the human genome), un-gapped alignment may fail to align a large number of reads. (2) Gapped alignment (e.g., BFAST,[13] Bowtie2,[14] BWA,[15] Novoalign, SHRiMP2,[16] SOAPaligner,[17] and SSAHA2[18]) allows mismatches, insertions, and deletions. Most gapped alignment tools implement the Smith-Waterman[19] or Needleman-Wunsch[20] algorithms. (3) Spliced alignment (e.g., GSNAP,[21] TopHat,[22] MapSplice,[23] OSA,[24] and SOAPsplice[25]) allows the long extension of gaps within the query reads. Biologically, such long gaps may represent intronic regions or inter-chromosomal splitting. Algorithmically, spliced alignment may be achieved by segmenting query reads into smaller sequences (e.g., 25 base pairs), mapping these smaller sequences, and then assembling mapped results for each read into a consensus result. Spliced alignment algorithms are often computationally more expensive than un-spliced algorithms. However, spliced mapping is necessary for applications that focus on identifying novel splice junctions using RNA-seq. Un-spliced mapping, including gapped and un-gapped, is sufficient for ChIP-seq data analysis.

Regarding mapping accuracy, it depends on mapping strategy. Uniquely mapped reads provide more definite information than multi-mapped reads. If a query read is mapped to multiple genomic loci due to insufficient read length, the ambiguous mapping happens and a mapping tool may randomly report one optimal mapping out of all mappings or report all optimal mappings. On the other hand, multi-mapped reads may benefit the downstream quantification algorithms in model training and expression estimation.

To improve the computational efficiency of sequence mapping, auxiliary data structures can be used to reduce the similarity search space such as to index either the reference genome or query reads using hash tables (BFAST, GSNAP, SHRiMP2, and SSAHA2 are representatives), or to index the reference genome using the Burrows-Wheeler transform with suffix/prefix arrays (Bowtie, Bowtie2, BWA, and SOAP2 are representatives).[26]

Expression Quantification

The second step of the RNA-seq bioinformatics pipeline is expression quantification of genes, transcripts, or other functional small RNAs. Because a read may map to multiple genomic loci, the accuracy of gene or transcript expression estimation depends on the ability of the quantification algorithm to resolve the ambiguities from the sequence-mapping step. In addition, a gene may have multiple alternatively spliced isoforms sharing a common set of exons, where a read mapped to the shared exons may belong to any one of the isoforms. Currently, the handling of these ambiguities involves building a probabilistic framework and then estimating gene/transcript expression using either the expectation-maximization (EM) algorithm or Bayesian inference.[27-29]

Quantification algorithms can be categorized into three groups: count-based, linear model-based, and Poisson model-based. [30] Table 2 lists common RNA-seq quantification tools, categorized in terms of the model, the estimation algorithm, and quantifiable targets. Count-

based quantifiers (e.g., ERANGE,[31] HTSeq,[32] NEUMA,[33] and ALEXA-Seq[34]) assign each read to its mapped location with a probability of one. Each quantifier implements a proprietary filtering criterion, and the expression profile is the accumulated read count on each targeted gene or transcript. Linear model-based quantifiers (e.g., rQuant[35] and IsoInfer[36]) assume that read counts are normally distributed, and least squares can be applied to infer expression estimates. Poisson model-based quantifiers (e.g., RSEM,[27] Cufflinks,[28] MISO,[29] and IsoEM[37]) probabilistically assign multi-mapped reads based on the assumption that reads from genomic loci follow the Poisson distribution. Because count-based quantifiers do not rely on a predefined model, they usually have lower computational complexity than the other model-based quantifiers. However, the expression estimates of count-based quantifiers might deviate from the truth because of the naïve way in which multi-mapped reads are handled.

Expression Normalization

The third step of the RNA-seq bioinformatics pipeline is normalization. Because of variations introduced in sequencing and bioinformatics processes, inter-sample comparison of RNA-seq expression estimates can only be done after normalization. Most normalization methods for RNA-seq are based on scaling, in which the gene or transcript expression of any biological sample is normalized by multiplying or dividing by a fixed scaling factor. Therefore, the fundamental challenge for RNA-seq expression normalization is to estimate a set of robust scaling factors for samples in the dataset. Table 3 lists commonly used RNA-seq normalization methods.

Several naïve methods such as RPM/FPM (reads/fragments per million mapped reads/fragments), median normalization, and upper-quartile normalization[38] are mathematically similar. RPM/FPM adjusts expression estimates of each sample by the total number of mapped reads/fragments in the sample. Median and upper-quartile normalizations use the median and upper quartile read/fragment counts, respectively, of each sample as the substitute for the total mapped reads/fragments. With the Illumina sequencing protocol, longer genes or transcripts tend to produce a larger number of sequence fragments. Thus, some methods such as RPKM/FPKM (reads/fragments per kilobase per million mapped reads/fragments)[31] and TPM (transcripts per million)[27] further adjust expression estimates by gene or transcript length, which in turn enables both inter- and intra-sample comparisons. However, there exist limitations for the aforementioned normalization methods. For the RPKM/FPKM, gene or transcript length cannot be precisely defined. Also some methods such as RPM/FPM and RPKM/FPKM are sensitive to “extreme” datasets that have a small number of highly differentially expressed genes. Therefore, a number of methods such as TMM (trimmed mean of M-values)[39] and RLE (relative log expression)[40] assume that most genes are not differentially expressed and use robust estimates of library size as the scaling factors. Dillies et al. systematically evaluated a few normalization methods and recommended TMM or RLE as the most robust method for most RNA-seq data.[41]

Differentially Expressed Gene Detection

The fourth key step in the identification of potential CVD biomarkers from RNA-seq data is the detection of DEGs between two groups of samples. RNA-seq expression estimates from

the two groups are first fit to a statistical distribution, followed by statistical hypothesis testing to determine whether the statistical distributions between the two groups are significantly different for a targeted gene. Soneson et al. and Rapaport et al. have conducted comprehensive quantitative evaluations for DEG detection methods.[42, 43] Thus, this section will mainly focus on qualitative categorization.

DEG detection methods can be nonparametric or parametric. Nonparametric methods such as SAMseq[44] and NOISeq[45] use resampling and counting techniques to avoid making assumptions about the underlying distribution of RNA-seq expression estimates. Data permutation is a common technique for estimating false discovery rates for nonparametric methods.

Parametric methods use Poisson-based models to fit RNA-seq read count data.[46] The Poisson distribution has the variance equals to the mean. However, overdispersion (i.e., when the variance is significantly greater than the mean) often occurs in RNA-seq data. Therefore, the negative binomial distribution, which is a two-parameter extension of the Poisson distribution, introduces an additional parameter to capture the high variability.[40] Selecting an appropriate statistical model is the key for a parametric DEG detection method. For example, DEGseq[47] applies the Poisson distribution to model RNA-seq read count data; edgeR,[48] baySeq,[49] and DESeq[40] use the negative binomial model to capture the overdispersion; Myrna[50] models the data as either the Gaussian distribution or the Poisson distribution; and Cuffdiff2[51] uses the beta negative binomial model to capture both overdispersion and uncertainty in the fragment count of a transcript. After constructing the statistical model, most parametric methods assess the significance level of each gene by using either p-values computed from the likelihood ratio test or the Fisher's exact test, or posterior probabilities estimated from the empirical Bayes method, while Cuffdiff2 assumes that RNA-seq data is normally distributed after a particular transformation and uses the t-test to determine the statistical significance of each DEG.

Peak Calling

For the ChIP-seq bioinformatics pipeline, the second step is peak calling. Sequence reads generated from ChIP-seq mostly originate from DNA sequences around targeted protein-binding regions. After mapping reads to the reference genome and identifying uniquely mapped reads, genomic loci that accumulate a large number of reads (i.e., peaks) indicate putative protein-binding regions. Peak-calling tools distinguish true peaks from background noise by (1) generating a signal profile along each chromosome, (2) defining a background noise model, (3) identifying candidate peak locations, and (4) assessing the significance of each candidate peak.[52] Peak-calling tools in earlier time quantify fold enrichment between samples of interest and expected background, and then apply the Poisson model to assess the significance of the enriched regions.[52] Recently developed peak-calling tools use the strand-dependent bimodality information and adopt a more realistic background model to capture local variations.[53]

Case Studies: RNA-seq and ChIP-seq Bioinformatics for CVD

In this section, we will demonstrate the utility of NGS bioinformatics for cardiovascular research by applying RNA-seq and ChIP-seq pipelines to publicly available CVD RNA-seq and ChIP-seq datasets downloaded from the NCBI Sequence Read Archive (SRA) repository (Figure 2). In each case study, we will illustrate the NGS bioinformatics solution for CVD research; evaluate the performance of the critical step that identifies CVD biomarkers; and discuss the remaining bioinformatics challenges.

Cardiovascular Transcriptomics Using RNA-seq

Dataset—The RNA-seq dataset (SRA accession: SRP009662) was acquired to investigate the effects of Ezh2 deletion on postnatal cardiac development, homeostasis, and gene expression.[54] The authors reported that the loss of Ezh2 gene in cardiac precursors would lead to cardiac hypertrophy and fibrosis. This dataset contains wild-type and Ezh2-deficient adult mouse right ventricle samples, each with two biological replicates. Each sample was sequenced with the Illumina HiSeq 2000 platform and contains around 30 million 2×50 bp read pairs.

Bioinformatics Pipeline and Performance Evaluation Metrics

The bioinformatics pipeline for identifying DEGs using RNA-seq data includes sequence mapping, expression quantification, expression normalization, and DEG detection (Figure 2). We use the same sequence mapper TopHat[22] and expression quantifier Cufflinks[28] with eight different DEG detection tools to construct eight pipelines. Each DEG detection tool implements an expression normalization method that optimizes its DEG detection performance. TopHat maps the four RNA-seq samples to the GRCm38/mm10 mouse genome[55] with the guidance of the RefSeq genome annotation.[56] Cufflinks quantifies gene/transcript expression in terms of raw read counts. The eight DEG detection tools (Figure 2, left table) include both nonparametric (e.g., SAMseq[44] and NOISeq[45]) and parametric methods (e.g., baySeq,[49] Cuffdiff2,[51] DESeq2,[40] DSS,[57] edgeR,[48] and Limma+Voom[58]) that represent a wide variety of prevalent and novel algorithms. The significant DEGs are identified with adjusted p-values less than 0.05. To evaluate the performance of the eight pipelines, we have designed five metrics:

1. The authors of the original paper[54] performed qRT-PCR validation for 16 cardiac hypertrophy- or fibrosis-related genes. Among 16 genes, only 12 were reported to be significantly differentially expressed. To assess the power of RNA-seq, we use the concordance of these 12 DEGs between qRT-PCR and eight RNA-seq pipelines as the first metric. For genes detected by less than four out of the eight pipelines, we further investigated those expression patterns to understand the cause.
2. To assess the biological relevance of DEGs, we use the ToppFun web-based tool in the ToppGene Suite[59] to annotate the functions of DEGs in terms of 114 significant Gene Ontology (GO) terms and four significant pathways with adjusted p-values less than 0.05. The GO terms and pathways associated with the 16 cardiac hypertrophy- or fibrosis-related

genes are defined as the ground-truth functional annotation (Figure 3, Panel B). We use the concordance between the pipeline-specific annotations and the ground-truth as the second metric.

3. To assess the reproducibility among various DEG detection tools, we compute the number of overlapping DEGs among the eight tools as the third metric.

4. To assess the expression profiles of DEGs, as shown in Equation (1), we used the ratio of the dominant read count (i.e., the largest read count of a DEG across all samples) to the total read count for any DEG g, and its distribution as the fourth metric.

$$R_{\text{dominance},g}=\frac{\text{Max}(A_{1,g},A_{2,g},B_{1,g},B_{2,g})}{\text{Sum}(A_{1,g},A_{2,g},B_{1,g},B_{2,g})}. \quad (1)$$

$A_{1,g}$, $A_{2,g}$, $B_{1,g}$, *and* $B_{2,g}$ are normalized read counts after adjusting the sequencing depth effect for samples A_1, A_2, B_1, *and* B_2 for any DEG g. [A_1, A_2] and [B_1, B_2] are biological replicates for the wild-type and Ezh2-deficient samples respectively. Given Equation (1), the range of $R_{dominance,g}$ is from 25% (i.e., $A_{1,g}$=$A_{2,g}$=$B_{1,g}$=$B_{2,g}$) to 100% (i.e., *Max*=*Sum*) with a few possible scenarios: (a) if a gene is not significantly differentially expressed and the variability between replicates is small, the normalized read counts $A_{1,g}$, $A_{2,g}$, $B_{1,g}$, *and* $B_{2,g}$ will only differ slightly from one another, and the $R_{dominance,g}$ will be close to 25%; (b) if a gene is highly differentially expressed and the variability between replicates is small, $R_{dominance,g}$ will be around 50%; and (c) if the variability between replicates is large, $R_{dominance,g}$ can be significantly greater than 50% (e.g., $R_{dominance,g}$ = 80% if [$A_{1,g}$, $A_{2,g}$, $B_{1,g}$, $B_{2,g}$] = [120, 30, 0, 0]).

5. To assess the capability of each tool for detecting highly-expressed and/or low-expressed DEGs, we calculate the mean read count of each DEG g from the normalized read counts (i.e., $A_{1,g}$, $A_{2,g}$, $B_{1,g}$, *and* $B_{2,g}$) and its distribution as the fifth metric.

Results and Discussion

We have evaluated the performance of the eight DEG detection tools by five metrics (Figure 3, Panels A-F). Using the 12 qRT-PCR validated DEGs, Panel A shows that nonparametric methods such as NOISeq and SAMseq, and parametric methods such as Cuffdiff2 and edgeR, were able to identify at least half of these 12 DEGs. In contrast, parametric methods, such as baySeq, DESeq2, DSS, and Limma+Voom, were able to identify only one or two of these 12 DEGs. Six genes were difficult to detect by RNA-seq-based methods (marked in red). The expression pattern of these difficult genes shows that they have either smaller fold changes (e.g., Tgfb3) or higher between-replicate variability (e.g., Actn3).

Panel B listed the top 20 GO terms and all four pathways from the ground-truth functional annotations, most of which were linked to the mechanisms of muscle contraction and heart development. Panel C summarized the concordance between the pipeline-specific annotations and the ground-truth in terms of the top 20 GO terms (ranking by p-values), all GO terms, and all pathways. DEGs detected by baySeq, DSS, and Limma+Voom were associated with zero GO terms and only a few pathways, which suggested that these tools detected DEGs with very diverse functions. DEGs detected by edgeR and Cuffdiff2 had more high-ranking functional annotations concordant with the ground-truth annotation. In contrast, DEGs detected by DESeq2, SAMseq, and NOISeq were linked to many GO terms and pathways that were biologically irrelevant to the original study, with no concordance appeared in the top 20 GO terms.

Panel D showed the number of DEGs supported by one, two, or all eight tools for each DEG detection method. baySeq, DSS, and Limma+Voom identified a fewer number of DEGs (i.e., 32, 23, and 12, respectively) that were highly reproducible among various tools (i.e., each DEG were supported by at least two other tools). Tools with more detected DEGs, such as SAMseq, NOISeq, and DESeq2, tended to have more pipeline-specific or unique DEGs. However, as discussed earlier, a higher number of DEGs did not necessarily lead to more biologically relevant results. Panel E demonstrated the distribution of $R_{dominance,g}$ for DEGs. Most DEGs had $R_{dominance,g}$ in the range of 25% to 60% following the scenarios (4a) and (4b) we have discussed in Section 3.1.2. Such observation indicated that most DEGs detected by RNA-seq pipelines did not have huge variability between biological replicates. DEGs with larger between-replicate variability resulted in $R_{dominance,g}$ greater than 60%. Such high variability can be the nature of biological replicates or biases introduced in the sequencing or bioinformatics processes. The nonparametric NOISeq method had the highest percentage of DEGs with $R_{dominance,g}$ greater than 60% since it identified many genes with very low read counts (e.g., $[A_{1,g}, A_{2,g}, B_{1,g}, B_{2,g}] = [1, 1, 0, 0]$). Thus, a small deviation in the read counts may have caused a huge variation in $R_{dominance,g}$. For parametric methods, higher $R_{dominance,g}$ indicated that the read counts may not follow a negative binomial distribution.[44] edgeR and Cuffdiff2 had a higher chance of detecting this type of genes as DEGs. Panel F showed the distribution of the mean read counts of DEGs. NOISeq had a bimodal distribution because of its tendency to identify some DEGs with very low read counts. The other seven tools shared a similar range of the mean read counts of DEGs, with baySeq slightly skewed to the left (i.e., lower mean read counts).

In summary, the original paper used RNA-seq to study the effect of Ezh2 deletion on gene expression profiles. It identified a set of DEGs relevant to cardiac tissue development and remodeling.[54] Our study examined the functions of DEGs detected by the eight RNA-seq pipelines, and edgeR and Cufflinks yielded the most functionally relevant DEGs. The nonparametric methods such as NOISeq and SAMseq identified many more DEGs than other tools, yet a large proportion of these DEGs may have been less reliable (e.g., DEGs with very low read counts) and irrelevant to the biology of the original study.

Remaining Bioinformatics Challenges

RNA-seq technology provides an opportunity to comprehensively study the transcriptome. While fixing sequence mapping and expression quantification steps and focusing on evaluating only DEG detection methods, we found that different tools generated very different DEG sets. Therefore, translating the computational findings into real clinical applications requires integrative biological interpretation and large-scale experimental validation. Capturing the full dynamics of the pipeline and forming a guideline of pipeline selection require a factorial experiment for studying the effect of each module in the pipeline. Finally, RNA-seq technology can be unreliable for low-expressing genes, but currently no standardized methods are capable of handling them properly. Thus, distinguishing true signals from noise for low-expressing genes remains a challenge.

Cardiovascular Epigenomics Using ChIP-seq

Dataset

The ChIP-seq dataset (SRA accession: SRP008658) investigated the genome-wide map of human heart enhancers with a pan-specific antibody that targets two closely-related transcriptional coactivator proteins, p300 and CBP (CREB-binding protein).[60] This dataset contains tissue samples from one fetal and one adult human heart. Each sample was sequenced with Illumina Genome Analyzer and contains around 27 million 36 bp single-ended reads.

Bioinformatics Pipeline and Performance Evaluation Metrics

The bioinformatics pipeline for identifying genome-wide protein-binding regions using ChIP-seq includes sequence mapping and peak calling (Figure 2). We use the same sequence mapper, Bowtie, and six different peak-calling tools to construct totally six pipelines. Bowtie[12] maps sequence reads (or sequence “tags” in ChIP-seq) to the GRCh37/hg19 human genome[61] and reports only uniquely mapped tags. The six peak-calling tools (Figure 2, right table), including SISSRs,[62] MACS,[63] FindPeaks,[64] SWEMBL,[65] SICER,[66] and F-Seq,[67] represent a wide variety of algorithms for determining statistically significant peaks. We run these tools using their default or recommended parameters with a p-value threshold of 10^{-3}. The identified peaks are putative protein-binding regions for p300 and CBP proteins.

To assess the performance of the six peak-calling tools, we have designed five metrics: (1) to visualize sequence-mapping and peak-calling information using the Integrative Genomics Viewer;[68] (2) to count the total number of peaks called by each tool; (3) to investigate the distribution of N_i, the normalized tags per peak, as defined in the Equation (2):

$$N_i = \frac{(\text{Number of Tags})_i}{(\text{Peak Length}/100)_i}; \quad (2)$$

(4) to compute the average length of peaks called by each tool, as defined in the Equation (3):

$$L_{\text{average}} = \frac{\sum_{i=1}^{N} \text{Length}(\text{Peak}_i)}{N}, \quad (3)$$

where N is the total number of peaks; and (5) to biologically validate the peaks by investigating the percentage of peaks that contain at least one p300 motif using FIMO (find individual motif occurrences)[69] with a p-value threshold of 10^{-4}. The input information for FIMO includes DNA sequences corresponding to these peaks and the position-specific scoring matrix for the p300 motif retrieved from the SwissRegulon Portal.[70]

Results and Discussion

We have investigated the performance of the six peak-calling tools by five metrics (Figure 4, Panels A-E). Visualizing by the Integrative Genomics Viewer, Panel A showed the peak regions called by the six tools with corresponding coverage information from the Bowtie alignment in the upstream region of the INPP5A gene (inositol polyphosphate-5-phsphatase, 40kDa). The selective nature of SISSRs and MACS resulted in sparse and short peaks. In contrast, FindPeaks and SWEMBL tended to call very long peaks with lengths over 10 kbp. Panel B demonstrated the number of peaks called by the six tools. SICER called the largest number of peaks, followed by FindPeaks and F-Seq. SICER failed to form longer peaks by merging nearby peaks, resulting in a relatively higher number of peaks. FindPeaks called two separate peaks even though two protein-binding regions were in close proximity; thus, FindPeaks also tended to call more peaks than the other tools. Panel C depicted the distribution of the number of tags per peak normalized by the peak length. Larger numbers indicated that the detected peaks were supported by more evidence. SISSRs, MACS, SWEMBL, and F-Seq exhibited a moderate to high number of tags per peak. In contrast, FindPeaks and SICER detected some peaks with a very low number of tags per peak. These peaks may not have been reliable because of limited evidence. Panel D showed the average length of peaks called by the six tools. Among them, SWEMBL had the longest average length, which may not have been a reasonable length for protein DNA-binding sites. SISSRs, MACS, and F-Seq exhibited the average peak length of less than 400 bp, which was close to the designed fragment length from the Illumina sequencing protocol.

Using FIMO, Panel E demonstrated the percentage of peaks that contained the p300 motif. MACS performed the best with 15% to 23% of the peaks containing the motif. SISSRs and F-Seq performed moderately well with their motif discovery rate ranging from 6% to 8%. Even though FindPeaks and SICER detected a significantly larger number of peaks than the others, only 2% to 3% of these peaks contained the p300 motif, exposing their relatively high false positive rate. Around 11% to 22% of peaks called by SWEMBL contained the p300 motif. However, despite such high performance, the peaks were not reliable since SWEMBL had extremely long peaks on average, which increased the probability of identifying the motif by chance alone.

In summary, the original study used ChIP-seq with the antibody that recognizes the enhancer-associated coactivator proteins p300 and CBP to annotate candidate heart enhancers that may regulate the expression of heart development-related genes in the human

genome. By examining the percentage of peak regions (i.e., candidate heart enhancer regions) that contained the p300 motif, MACS achieved the highest motif discovery rate among the six tools, which suggested that MACS identified more biologically relevant peaks than the others. In addition, MACS's peaks had the second highest tag coverage and the reasonable average peak length. In contrast, SICER identified peaks with the lowest motif discovery rate and very low tag coverage.

Remaining Bioinformatics Challenges

Similar to the case for RNA-seq, ChIP-seq requires a factorial experiment for studying the effect of either sequence-mapping or peak-calling step. Most peak-calling tools need control samples for building background models essential for conducting statistical tests. These background models can be local or global. The global model is easier to build but lacks the consideration of local biases. Accurately identifying peaks requires an adaptive background signal model that can dynamically change parameters to accommodate local variations and different ChIP-seq experiments.

The Opportunities of NGS for CVD

The prevention or treatment of CVDs can benefit from personalized care that tailors clinical decisions on a patient-by-patient basis. Most common CVDs such as atherosclerosis and coronary artery disease have complex phenotypes and are affected by both genetic and environmental factors.[71] Recently, omic biomarkers (e.g., genomic, transcriptomic, and epigenomic) have emerged as a complement to traditional CVD risk factors. Facilitated by NGS bioinformatics, one distinctive feature of the NGS technology is its capability to identify a variety of omic biomarkers. As shown in the two case studies in Section 3, the benefits of NGS bioinformatics are (1) it can comprehensively detect DEGs or peaks in the entire genome; (2) it can define the boundaries of peaks using ChIP-seq data at base-pair resolution; and (3) it can discover previously unknown DEGs or peaks as candidate omic biomarkers. Currently, most research uses a single omic data modality captured by NGS to study only one aspect of biological mechanisms of diseases. However, for complex diseases such as CVDs, candidate biomarkers based on a single omic data modality suffer from the reproducibility issue.[72] To address this issue, we need to research novel NGS bioinformatics methods that not only can integrate multiple omic data modalities (e.g., genetic variations, transcriptional regulation, and epigenetic modifications), but also can associate heterogeneous omic information with CVD phenotypes. Associations (e.g., correlation and causality) resulted from NGS bioinformatics present a great opportunity for researchers to obtain further insights on disease mechanisms and to improve risk predictions for complex CVDs.

Acknowledgments

Funding Sources: This work was supported in part by grants from the National Institutes of Health (NHLBI 5U01HL080711, U54CA119338, 1RC2CA148265), the Georgia Cancer Coalition (Distinguished Cancer Scholar Award to Professor May D. Wang), the Children's Healthcare of Atlanta, Microsoft Research, and the Hewlett-Packard Company.

References

1. Wilson PWF, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. Circulation. 1998; 97:1837–1847. [PubMed: 9603539]
2. Berenson GS, Srinivasan SR, Bao WH, Newman WP, Tracy RE, Wattigney WA, et al. Association between multiple cardiovascular risk factors and atherosclerosis in children and young adults. N Engl J Med. 1998; 338:1650–1656. [PubMed: 9614255]
3. Thanassoulis G, Vasan RS. Genetic cardiovascular risk prediction will we get there? Circulation. 2010; 122:2323–2334. [PubMed: 21147729]
4. Arnett DK, Baird AE, Barkley RA, Basson CT, Boerwinkle E, Ganesh SK, et al. Relevance of genetics and genomics for prevention and treatment of cardiovascular disease: A scientific statement from the american heart association council on epidemiology and prevention, the stroke council, and the functional genomics and translational biology interdisciplinary working group. Circulation. 2007; 115:2878–2901. [PubMed: 17515457]
5. Schnabel RB, Baccarelli A, Lin H, Ellinor PT, Benjamin EJ. Next steps in cardiovascular disease genomic research--sequencing, epigenetics, and transcriptomics. Clin Chem. 2012; 58:113–126. [PubMed: 22100807]
6. Ware JS, Roberts AM, Cook SA. Next generation sequencing for clinical diagnostics and personalised medicine: Implications for the next generation cardiologist. Heart. 2012; 98:276–281. [PubMed: 22128206]
7. Song HK, Hong SE, Kim T, Kim DH. Deep rna sequencing reveals novel cardiac transcriptomic signatures for physiological and pathological hypertrophy. PLoS One. 2012; 7
8. Guo W, Schafer S, Greaser ML, Radke MH, Liss M, Govindarajan T, et al. Rbm20, a gene for hereditary cardiomyopathy, regulates titin splicing. Nat Med. 2012; 18:766–773. [PubMed: 22466703]
9. Schlesinger J, Schueler M, Grunert M, Fischer JJ, Zhang Q, Krueger T, et al. The cardiac transcription network modulated by gata4, mef2a, nkx2.5, srf, histone modifications, and micrornas. Plos Genet. 2011; 7
10. He A, Kong SW, Ma Q, Pu WT. Co-occupancy by multiple cardiac transcription factors identifies transcriptional enhancers active in heart. Proc Natl Acad Sci U S A. 2011; 108:5632–5637. [PubMed: 21415370]
11. Blow MJ, McCulley DJ, Li Z, Zhang T, Akiyama JA, Holt A, et al. Chip-seq identification of weakly conserved heart enhancers. Nat Genet. 2010; 42:806–810. [PubMed: 20729851]
12. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009; 10
13. Homer N, Merriman B, Nelson SF. Bfast: An alignment tool for large scale genome resequencing. PLoS One. 2009; 4:A95–A106.
14. Langmead B, Salzberg SL. Fast gapped-read alignment with bowtie 2. Nat Methods. 2012; 9:357–U354. [PubMed: 22388286]
15. Li H, Durbin R. Fast and accurate short read alignment with burrows-wheeler transform. Bioinformatics. 2009; 25:1754–1760. [PubMed: 19451168]
16. David M, Dzamba M, Lister D, Ilie L, Brudno M. Shrimp2: Sensitive yet practical short read mapping. Bioinformatics. 2011; 27:1011–1012. [PubMed: 21278192]
17. Li RQ, Yu C, Li YR, Lam TW, Yiu SM, Kristiansen K, et al. Soap2: An improved ultrafast tool for short read alignment. Bioinformatics. 2009; 25:1966–1967. [PubMed: 19497933]
18. Ning ZM, Cox AJ, Mullikin JC. Ssaha: A fast search method for large DNA databases. Genome Res. 2001; 11:1725–1729. [PubMed: 11591649]
19. Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol. 1981; 147:195–197. [PubMed: 7265238]
20. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol. 1970; 48:443–453. [PubMed: 5420325]

21. Wu TD, Nacu S. Fast and snp-tolerant detection of complex variants and splicing in short reads. Bioinformatics. 2010; 26:873–881. [PubMed: 20147302]
22. Trapnell C, Pachter L, Salzberg SL. Tophat: Discovering splice junctions with rna-seq. Bioinformatics. 2009; 25:1105–1111. [PubMed: 19289445]
23. Wang K, Singh D, Zeng Z, Coleman SJ, Huang Y, Savich GL, et al. Mapsplice: Accurate mapping of rna-seq reads for splice junction discovery. Nucleic Acids Res. 2010; 38:e178. [PubMed: 20802226]
24. Hu J, Ge H, Newman M, Liu K. Osa: A fast and accurate alignment tool for rna-seq. Bioinformatics. 2012; 28:1933–1934. [PubMed: 22592379]
25. Huang S, Zhang J, Li R, Zhang W, He Z, Lam TW, et al. Soapsplice: Genome-wide ab initio detection of splice junctions from rna-seq data. Front Genet. 2011; 2:46. [PubMed: 22303342]
26. Li H, Homer N. A survey of sequence alignment algorithms for next-generation sequencing. Brief Bioinform. 2010; 11:473–483. [PubMed: 20460430]
27. Li B, Dewey CN. Rsem: Accurate transcript quantification from rna-seq data with or without a reference genome. BMC Bioinformatics. 2011; 12:323. [PubMed: 21816040]
28. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010; 28:511–515. [PubMed: 20436464]
29. Katz Y, Wang ET, Airoldi EM, Burge CB. Analysis and design of rna sequencing experiments for identifying isoform regulation. Nat Methods. 2010; 7:1009–1015. [PubMed: 21057496]
30. Pachter L. Models for transcript quantification from rna-seq. arXiv:1104.3889v2. 2011
31. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by rna-seq. Nat Methods. 2008; 5:621–628. [PubMed: 18516045]
32. Anders S. Htseq: Analysing high-throughput sequencing data with python. 2010
33. Lee S, Seo CH, Lim B, Yang JO, Oh J, Kim M, et al. Accurate quantification of transcriptome from rna-seq data by effective length normalization. Nucleic Acids Res. 2011; 39:e9. [PubMed: 21059678]
34. Griffith M, Griffith OL, Mwenifumbo J, Goya R, Morrissy AS, Morin RD, et al. Alternative expression analysis by rna sequencing. Nat Methods. 2010; 7:843–847. [PubMed: 20835245]
35. Bohnert R, Ratsch G. Rquant.Web: A tool for rna-seq-based transcript quantitation. Nucleic Acids Res. 2010; 38:W348–351. [PubMed: 20551130]
36. Feng J, Li W, Jiang T. Inference of isoforms from short sequence reads. J Comput Biol. 2011; 18:305–321. [PubMed: 21385036]
37. Nicolae M, Mangul S, Mandoiu II, Zelikovsky A. Estimation of alternative splicing isoform frequencies from rna-seq data. Algorithms Mol Biol. 2011; 6:9. [PubMed: 21504602]
38. Bullard JH, Purdom E, Hansen KD, Dudoit S. Evaluation of statistical methods for normalization and differential expression in mrna-seq experiments. BMC Bioinformatics. 2010; 11:94. [PubMed: 20167110]
39. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of rna-seq data. Genome Biol. 2010; 11:R25. [PubMed: 20196867]
40. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010; 11:R106. [PubMed: 20979621]
41. Dillies MA, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, et al. A comprehensive evaluation of normalization methods for illumina high-throughput rna sequencing data analysis. Brief Bioinform. 2012
42. Soneson C, Delorenzi M. A comparison of methods for differential expression analysis of rna-seq data. BMC Bioinformatics. 2013; 14:91. [PubMed: 23497356]
43. Rapaport F, Khanin R, Liang Y, Pirun M, Krek A, Zumbo P, et al. Comprehensive evaluation of differential gene expression analysis methods for rna-seq data. Genome Biol. 2013; 14:R95. [PubMed: 24020486]
44. Li J, Tibshirani R. Finding consistent patterns: A nonparametric approach for identifying differential expression in rna-seq data. Stat Methods Med Res. 2013; 22:519–536. [PubMed: 22127579]

45. Tarazona S, Garcia-Alcalde F, Dopazo J, Ferrer A, Conesa A. Differential expression in rna-seq: A matter of depth. Genome Res. 2011; 21:2213–2223. [PubMed: 21903743]
46. Chen Z, Liu J, Ng HK, Nadarajah S, Kaufman HL, Yang JY, et al. Statistical methods on detecting differentially expressed genes for rna-seq data. BMC Syst Biol. 2011; 5(3):S1.
47. Wang L, Feng Z, Wang X, Zhang X. Degseq: An r package for identifying differentially expressed genes from rna-seq data. Bioinformatics. 2010; 26:136–138. [PubMed: 19855105]
48. Robinson MD, McCarthy DJ, Smyth GK. Edger: A bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010; 26:139–140. [PubMed: 19910308]
49. Hardcastle TJ, Kelly KA. Bayseq: Empirical bayesian methods for identifying differential expression in sequence count data. BMC Bioinformatics. 2010; 11:422. [PubMed: 20698981]
50. Langmead B, Hansen KD, Leek JT. Cloud-scale rna-sequencing differential expression analysis with myrna. Genome Biol. 2010; 11
51. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L. Differential analysis of gene regulation at transcript resolution with rna-seq. Nat Biotechnol. 2013; 31:46–53. [PubMed: 23222703]
52. Pepke S, Wold B, Mortazavi A. Computation for chip-seq and rna-seq studies. Nat Methods. 2009; 6:S22–32. [PubMed: 19844228]
53. Wilbanks EG, Facciotti MT. Evaluation of algorithm performance in chip-seq peak detection. PLoS One. 2010; 5:e11471. [PubMed: 20628599]
54. Delgado-Olguin P, Huang Y, Li X, Christodoulou D, Seidman CE, Seidman JG, et al. Epigenetic repression of cardiac progenitor gene expression by ezh2 is required for postnatal cardiac homeostasis. Nat Genet. 2012; 44:343–347. [PubMed: 22267199]
55. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, et al. Mouse Genome Sequencing C. Initial sequencing and comparative analysis of the mouse genome. Nature. 2002; 420:520–562. [PubMed: 12466850]
56. Pruitt KD, Tatusova T, Brown GR, Maglott DR. Ncbi reference sequences (refseq): Current status, new features and genome annotation policy. Nucleic Acids Res. 2012; 40:D130–135. [PubMed: 22121212]
57. Wu H, Wang C, Wu Z. A new shrinkage estimator for dispersion improves differential expression detection in rna-seq data. Biostatistics. 2013; 14:232–243. [PubMed: 23001152]
58. Law CW, Chen Y, Shi W, Smyth GK. Voom: Precision weights unlock linear model analysis tools for rna-seq read counts. Genome Biol. 2014; 15:R29. [PubMed: 24485249]
59. Chen J, Bardes EE, Aronow BJ, Jegga AG. Toppgene suite for gene list enrichment analysis and candidate gene prioritization. Nucleic Acids Res. 2009; 37:W305–311. [PubMed: 19465376]
60. May D, Blow MJ, Kaplan T, McCulley DJ, Jensen BC, Akiyama JA, et al. Large-scale discovery of enhancers from human heart tissue. Nat Genet. 2012; 44:89–93. [PubMed: 22138689]
61. International Human Genome Sequencing C. Finishing the euchromatic sequence of the human genome. Nature. 2004; 431:931–945. [PubMed: 15496913]
62. Jothi R, Cuddapah S, Barski A, Cui K, Zhao K. Genome-wide identification of in vivo protein-DNA binding sites from chip-seq data. Nucleic Acids Res. 2008; 36:5221–5231. [PubMed: 18684996]
63. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of chip-seq (macs). Genome Biol. 2008; 9:R137. [PubMed: 18798982]
64. Fejes AP, Robertson G, Bilenky M, Varhol R, Bainbridge M, Jones SJ. Findpeaks 3.1: A tool for identifying areas of enrichment from massively parallel short-read sequencing technology. Bioinformatics. 2008; 24:1729–1730. [PubMed: 18599518]
65. Wilder S. Swembl: A generic peak-calling program. 2010
66. Zang C, Schones DE, Zeng C, Cui K, Zhao K, Peng W. A clustering approach for identification of enriched domains from histone modification chip-seq data. Bioinformatics. 2009; 25:1952–1958. [PubMed: 19505939]
67. Boyle AP, Guinney J, Crawford GE, Furey TS. F-seq: A feature density estimator for high-throughput sequence tags. Bioinformatics. 2008; 24:2537–2538. [PubMed: 18784119]

68. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nat Biotechnol. 2011; 29:24–26. [PubMed: 21221095]

69. Grant CE, Bailey TL, Noble WS. Fimo: Scanning for occurrences of a given motif. Bioinformatics. 2011; 27:1017–1018. [PubMed: 21330290]

70. Pachkov M, Erb I, Molina N, van Nimwegen E. Swissregulon: A database of genome-wide annotations of regulatory sites. Nucleic Acids Res. 2007; 35:D127–131. [PubMed: 17130146]

71. Epstein JA, Rader DJ, Parmacek MS. Perspective: Cardiovascular disease in the postgenomic era--lessons learned and challenges ahead. Endocrinology. 2002; 143:2045–2050. [PubMed: 12021168]

72. Faita F, Vecoli C, Foffa I, Andreassi MG. Next generation sequencing in cardiovascular diseases. World J Cardiol. 2012; 4:288–295. [PubMed: 23110245]

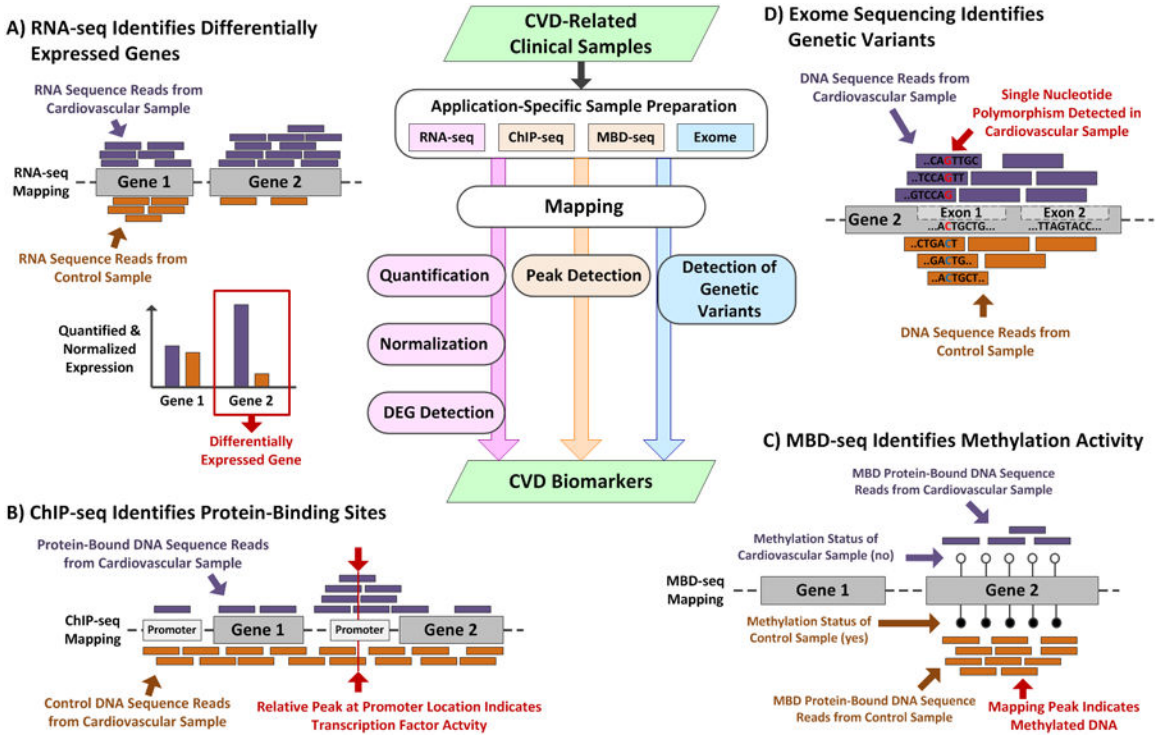

Figure 1.

Next-Generation Sequencing Facilitates the Identification of Omic Biomarkers for Cardiovascular Diseases. (A) RNA-seq detects differentially expressed genes by comparing gene expression profiles of CVD samples to those of control samples. (B) ChIP-seq identifies transcription factor activity by detecting peaks formed by mapping DNA sequence reads that bind to transcription factor proteins. Transcription factor activity correlates with gene expression. (C) Exome sequencing detects genetic variants such as SNPs that may correlate with CVD phenotypes. (D) MBD-seq is similar to ChIP-seq, but identifies regions of DNA methylation, which can affect gene expression.

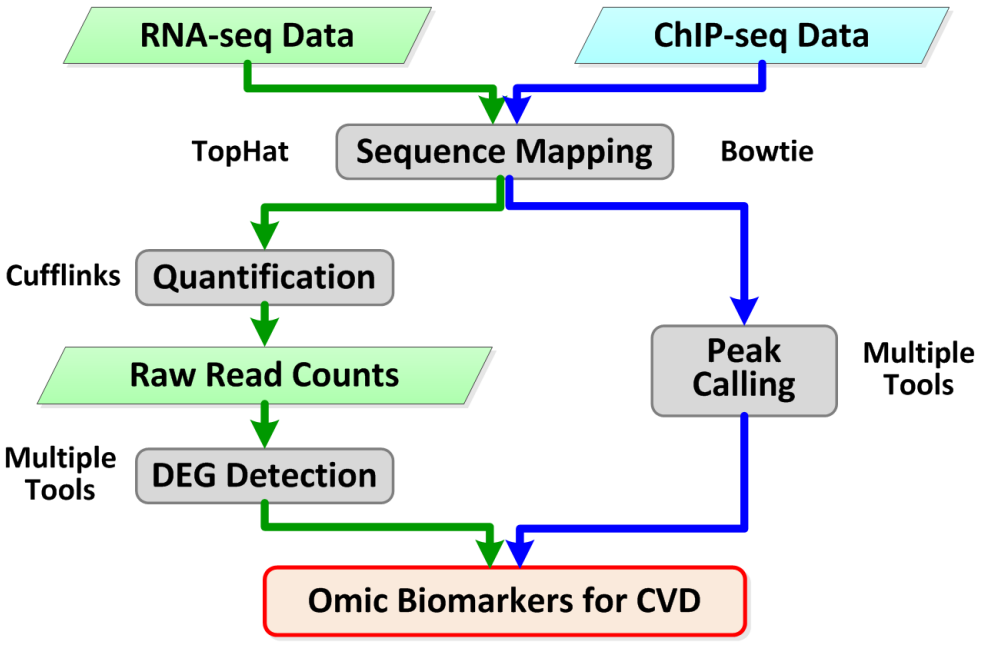

Figure 2.

Bioinformatics Pipelines and Tools for RNA-seq and ChIP-seq Case Studies. Green arrows indicate the pipeline for RNA-seq data and blue arrows represent the pipeline for ChIP-seq data. DEG detection and peak-calling tools are listed following alphabetical order.

(A)

qRT-PCR Validated DEGs	Tools (# DEGs)	baySeq (32)	Cuffdiff2 (155)	DESeq2 (514)	DSS (23)	edgeR (243)	Limma+ Voom (12)	NOISeq (792)	SAMseq (2401)	Expression Pattern			
										A1	A2	B1	B2
	Nppa		V			V	V	V	V	81596	57346	9593	3985
	Nppb		V			V		V	V	9201	6232	2815	1161
	Myh7		V			V		V	V	30488	17510	1822	7235
	Tgfb3									675	393	523	267
	Postn		V	V		V		V	V	1717	1252	320	539
	Spp1								V	33	40	4	21
	Six1	V		V	V	V		V	V	116	108	14	3
	Eya1								V	46	33	11	19
	Myh8							V	V	4	3	0	0
	Myl1		V			V		V	V	606	1536	199	128
	Myl4		V						V	631	500	332	35
	Actn3					V		V	V	94	4	2	0
	Total	1/12	6/12	2/12	1/12	7/12	1/12	8/12	11/12				

(B)

	baySeq (32)	Cuffdiff2 (155)	DESeq2 (514)	DSS (23)	edgeR (243)	Limma+ Voom (12)	NOISeq (792)	SAMseq (2401)
Number of Significant GO Terms*	0	107	93	0	75	0	171	406
Number of Concordant GO Terms between Ground Truth and Each Tool (Top 20)	0	8	0	0	12	0	0	0
Number of Concordant GO Terms between Ground Truth and Each Tool (Top 50)	0	11	1	0	20	0	7	3
Number of Significant Pathways*	0	18	20	2	17	0	28	74
Number of Concordant Pathways between Ground Truth and Each Tool **	0	2	0	0	3	0	3	3

* False Discovery Rate = 0.05
** Ground truth contains 4 significant pathways

(C)

GO Term	Adjusted P-Value
GO:0030049 Muscle filament sliding	1.33E-14
GO:0033275 Actin-myosin filament sliding	1.33E-14
GO:0070252 Actin-mediated cell contraction	9.77E-13
GO:0008307 Structural constituent of muscle	2.90E-12
GO:0030048 Actin filament-based movement	5.89E-11
GO:0003012 Muscle system process	5.89E-11
GO:0044449 Contractile fiber part	5.37E-10
GO:0005859 Muscle myosin complex	5.37E-10
GO:0043292 Contractile fiber	7.10E-10
GO:0016460 Myosin II complex	1.12E-09
GO:0015629 Actin cytoskeleton	1.12E-09
GO:0016459 Myosin complex	1.12E-09
GO:0006936 Muscle contraction	1.20E-09
GO:0030016 Myofibril	1.08E-08
GO:0007507 Heart development	2.54E-07
GO:0030017 Sarcomere	2.65E-07
GO:0006941 Striated muscle contraction	2.81E-07
GO:0003007 Heart morphogenesis	9.56E-07
GO:0003779 Actin binding	4.27E-06
GO:0061061 Muscle structure development	4.97E-06

Pathway	Adjusted P-Value
Reactome: Genes involved in Striated Muscle Contraction	5.33E-12
WikiPathways: Striated Muscle Contraction	1.27E-11
Reactome: Genes involved in Muscle contraction	7.72E-11
BioCarta: ALK in cardiac myocytes	2.85E-05

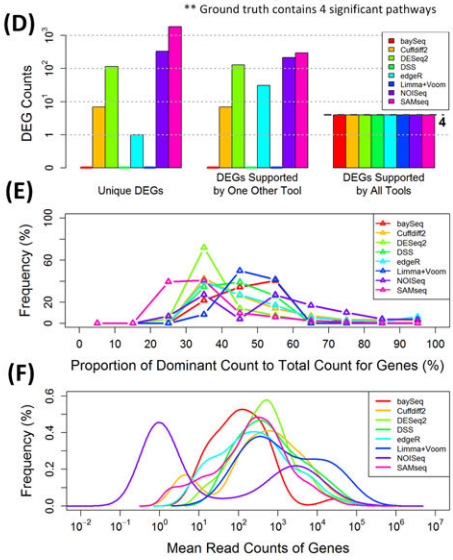

Figure 3.

Biological Interpretation and Quantitative Assessment of Various DEG Detection Tools Using RNA-seq Data. (A) Checkmarks indicated concordant DEGs between qRT-PCR validation conducted by the original study and the results from our eight DEG detection tools. Genes marked in red were identified by less than four out of the eight tools. (B) The top 20 significant GO terms and all four significant pathways of the *ground truth functional annotation*, which was established by annotating the 16 qRT-PCR validated genes. (C) The number of concordant GO terms and pathways between ground-truth and pipeline-specific annotations. (D) The number of DEGs supported by one, two, or all eight tools for each DEG detection pipeline. (E) The distribution of the ratio of dominant read counts to total read counts of all DEGs for each DEG detection pipeline. (F) The distribution of the mean read counts of all DEGs for each DEG detection pipeline.

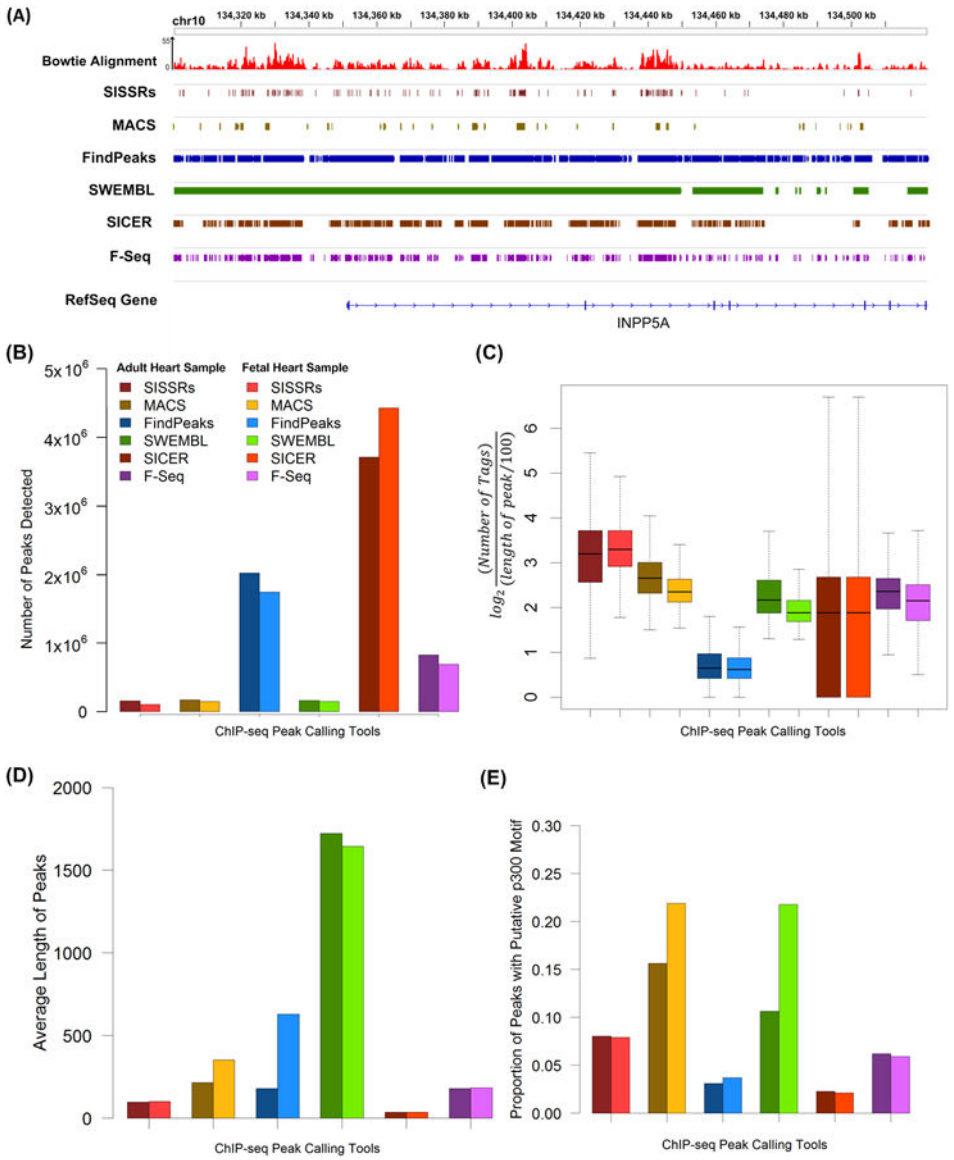

Figure 4.

Qualitative and Quantitative Assessment of Various Peak-Calling Tools Using ChIP-seq Data. (A) IGV visualized peaks called by the six tools in the upstream region of the INPP5A gene using the adult heart sample. The black histogram on the top represented the coverage of Bowtie alignment in the same region. (B) The number of peaks detected by each peak-calling pipeline. (C) The distribution of the number of tags per peak normalized by the peak length for each peak-calling pipeline. (D) The average length of peaks for each peak-calling pipeline. (E) The percentage of peaks that contains at least one p300 motif identified by FIMO with a p-value threshold of 10^{-4} for each peak-calling pipeline.

Table 1

Summary of RNA-seq and ChIP-seq Sequence-Mapping Tools

Sequence-Mapping Tool	Mapping Strategy and Usage	Algorithmic Notes
BFAST (BLAT-like fast accurate search tool)	Un-spliced mapping to transcriptome or genome	Hash table, Smith-Waterman local alignment
Bowtie		Burrows-Wheeler transform and FM-index
Bowtie2		Burrows-Wheeler transform, FM-index-assisted seed alignment, dynamic programming
BWA (Burrow-Wheeler aligner)		Burrows-Wheeler transform
Novoalign		*Commercial software, algorithm un-published
SHRiMP2 (short read mapping package, version 2)		Multiple spaced-seed indexing, Smith-Waterman local alignment
SOAPaligner (short oligonucleotide analysis package aligner)		Bi-directional Burrows-Wheeler transform
SSAHA2 (sequence search and alignment by hashing algorithm, version 2)		Hash table
GSNAP (genomic short-read nucleotide alignment program)	Spliced mapping to genome & un-spliced mapping to transcriptome or genome	Minimal sampling strategy, oligomer chaining for approximate alignment, sandwich dynamic programming
MapSplice	Spliced mapping to genome	Uses Bowtie for alignment, segmented mapping
OSA (Omicsoft sequence aligner)		Two-stage transcriptome and genome alignment, segmented mapping
SoapSpliceSOAPsplice (short oligonucleotide analysis package for splice junction detection)		Burrows-Wheeler transform, segmented mapping
TopHat		Uses Bowtie or Bowtie2 for alignment, segmented mapping

Table 2

Summary of RNA-seq Expression Quantification Tools

Quantification Tool	Mathematical Model	Estimation	Gene/Isoform
ALEXA-Seq (alternative expression analysis by sequencing)	Count-based model	Average coverage of mapped reads	Yes/Yes
ERANGE (enhanced read analysis of gene expression)		Accumulated counts, read assigns proportionally to expression level	Yes/No
HTSeq (analyzing high-throughput sequencing data with Python)		Accumulated counts, read assigns with probability 1	Yes/No
NEUMA (normalization by expected uniquely mappable area)		Accumulated counts of informative reads	Yes/Yes
IsoInfer (inference of isoforms from short sequence reads)	Linear model	Maximum likelihood estimation from convex quadratic programming	Yes/Yes
rQuant (transcript quantification with RNA-seq data)		Minimize read coverage deviation with quadratic programming	Yes/Yes
Cufflinks	Poisson model	Maximize likelihood with the maximum a posteriori estimates using Bayesian inference	Yes/Yes
IsoEM (isoform quantification by expectation maximization)		Maximum likelihood estimation with EM algorithm	Yes/Yes
MISO (mixture of isoforms)		Posterior mean estimates using Bayesian inference	Yes/Yes
RSEM (RNA-seq by expectation maximization)		Maximum likelihood estimation with EM algorithm	Yes/Yes

Table 3
Summary of RNA-seq Expression Normalization Methods

Normalization Method	Description
Median	Scaling by median of all counts
Quantile	Matching distributions of counts
RLE (relative log expression)	Scaling by median ratio to median library
RPKM/FPKM (reads/fragments per kilobase per million mapped reads/fragments)	Scaling by library size and gene/transcript length
RPM/FPM (reads/fragments per million mapped reads/fragments)	Scaling by library size
TMM (trimmed mean of M-values)	Scaling by estimate of relative RNA production
TPM (transcripts per million)	Scaling by mean length of expressed genes/transcripts
Upper Quartile	Scaling by upper quartile of all counts